AFHRL-TP-89-67

(2)

AIR FORCE

HUMAN RESOURCES

LABORATORY

ADDING A DIMENSION: TIME AS A FACTOR IN THE GENERALIZABILITY OF PREDICTIVE RELATIONSHIPS

Charles L. Hulin
Rebecca A. Henry
Sharon L. Noon

University of Illinois
Department of Psychology
603 East Daniel Street
Champaign, Illinois 61820

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

January 1990
Interim Technical Paper

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

WILLIAM E. ALLEY, Technical Director
Manpower and Personnel Division

DANIEL L. LEIGHTON, Colonel, USAF
Chief, Manpower and Personnel Division

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	January 1990	Interim

4. TITLE AND SUBTITLE

Adding a Dimension: Time as a Factor in the Generalizability of Predictive Relationships

5. FUNDING NUMBERS

C - F33615-87-C-0014
PE - 62703F
PR - 7719
TA - 18
WU - 55

6. AUTHOR(S)

Charles L. Hulin
Rebecca A. Henry
Sharon L. Noon

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Illinois
Department of Psychology
603 East Daniel Street
Champaign, Illinois 61820

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Manpower and Personnel Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235-5601

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

AFHRL-TP-89-67

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

Analyses of trends in predictive validity coefficients across time and repeated performance assessment ... ws highly significant and consistent trends in validities as a function of time and/or interpolated practice. Commonly used ability measures show decreasing predictive validities for the prediction of temporally more remote performance assessments. Within study corrections for differential restrictions of range and attenuation due to unreliability across the different performance assessments increased the negative slopes of the regressions of predictive validity on time or ordinal position of performance assessment. The median validity decrement from initial to final performance assessment, corrected for differential range restriction, attenuation, and within study sampling fluctuations was -.29. The mean of the trimmed distribution of corrected validity decrements, after eliminating the two most extreme cases, was -.45. The average within study correlation between predictive validity and time or ordinal position of performance assessment was 0.80. A similar analysis of stability coefficients of time period-by-time period or trial-by-trial performance assessment correlations revealed very similar albeit slightly more consistent findings. Theoretical explanations stressing the dynamic nature of human abilities, the changing nature of abilities required for task performance, and social competition factors are discussed as reasons for the predictive validity decrements.

14. SUBJECT TERMS

validity decay
meta-analysis
predictive validity

15. NUMBER OF PAGES

38

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

AFHRL Technical Paper 89-67 January 1990

ADDING A DIMENSION: TIME AS A FACTOR IN THE GENERALIZABILITY OF PREDICTIVE RELATIONSHIPS

Charles L. Hulin
Rebecca A. Henry
Sharon L. Noon

University of Illinois
Department of Psychology
603 East Daniel Street
Champaign, Illinois 61820

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

Reviewed by

Thomas R. Carretta
Aircrew Selection and Classification Function

Submitted for publication by

Joseph L. Weeks
Chief, Cognitive Skills Assessment Branch

SUMMARY

An analysis of trends in predictive validity coefficients across time and repeated performance assessments shows highly significant and consistent trends in validities as a function of time and/or interpolated practice. Commonly used ability measures show decreasing predictive validities for the prediction of temporally more remote performance assessments. Within study corrections for differential restrictions of range and attenuation due to unreliability across the different performance assessments increased the negative slopes of the regressions of predictive validity on time or ordinal position of performance assessment. The median validity decrement from initial to final performance assessment, corrected for differential range restriction, attenuation, and within study sampling fluctuations was -.29. The mean of the trimmed distribution of corrected validity decrements, after eliminating the two most extreme cases, was -.45. The average within study correlation between predictive validity and time or ordinal position of performance assessment was -.80. A similar analysis of stability coefficients of time period-by-time period or trial-by-trial performance assessment correlations revealed very similar albeit slightly more consistent findings. Theoretical explanations stressing the dynamic nature of human abilities, the changing nature of abilities required for task performance, and social competition factors are discussed as reasons for the predictive validity decrements.

PREFACE

This technical paper contains a meta-analysis of empirical articles containing data relevant to questions about temporal declines in predictive validities and declines in relationships between assessments of skilled performance. This meta-analysis establishes the framework within which ongoing studies being conducted at the University of Illinois of the validity of explanations for the observed declines can be interpreted. This meta-analysis confirms the generality of the phenomenon across a wide variety of skilled and cognitive performance areas. With one possible exception of performance in law school, there appear to be no performance areas immune to these declines in predictive validities and performance stabilities across time and repeated trials. The magnitude of the decline varies with initial predictive validity and the length of the study, but its magnitude is sufficient, over time, to raise serious questions about the benefits of using selection tests. Current investigations of reasons for these validity and stability declines may provide theoretical explanations for these powerful effects.

The authors thank James Austin, Kathy Hanisch, Lloyd Humphreys, and Mary Roznowski for comments on earlier drafts of this manuscript. Their comments improved and strengthened the analyses and interpretation of the results. This study was supported in part by Contract #F33615-87-C-0014 from Brooks AFB, Texas. The views of the authors do not necessarily represent those of the Air Force Human Resources Laboratory.

Requests for reprints may be sent to Dr. Charles L. Hulin, Department of Psychology, 603 East Daniel, Champaign, IL 61820.

TABLE OF CONTENTS

LIST OF TABLES

ADDING A DIMENSION: TIME AS A FACTOR IN THE GENERALIZABILITY OF PREDICTIVE RELATIONSHIPS

I. INTRODUCTION

Studies of the predictions of future performance from current abilities have typically ignored the time facet in prediction equations. With the exception of Alvares and Hulin (1972), Fleishman (1960), Humphreys (1968), and Humphreys and Taber (1973), the goals of most investigators have been to establish generalizability across populations of individuals, abilities (used as predictors), tasks, and situations. Most analyses of the generalizability of predictive relations have examined whether variance in predictive validities across elements of these four facets or populations can be attributed to statistical artifacts or to real differences in predictive relationships (Hunter, Schmidt, & Jackson, 1982; Schmidt & Hunter, 1977; Schmidt, Hunter, & Caplan, 1981).

A narrative review by Henry and Hulin (1987) of the literature relevant to the stability of predictive validities across time suggested that most empirical predictive validities were less stable than has commonly been assumed and that the instability was general across content areas (Henry & Hulin, 1987). They reported that temporally decreasing predictive validities have been found in most areas of skilled performance. Psychomotor skills such as discriminant reaction time (Fleishman & Hempel, 1954, 1955), two dimensional tracking (Dunham, 1974), rotary pursuit (Fleishman, 1960), two-handed coordination (Fleishman & Rich, 1963), and student pilot performance during training (Alvares & Hulin, 1973) were typically found to have decreasing predictive validities or decreasing intertrial correlations. Studies of the predictive validities for academic performance in college (Humphreys, 1968; Humphreys & Taber, 1973) and graduate school (Lin & Humphreys, 1977) also reported systematically changing predictive validities when evaluated against performance assessed at different stages of learning or performance. The time periods examined in such studies have ranged from one- or 2-hour experiments (Dunham, 1974; Fleishman, 1960), to performance across 15 weeks of flight training (Alvares & Hulin, 1973), to performance of engineers across 20 years (Brenner & Lockwood, 1965), to performance of scientists across five decades (Dennis, 1954, 1956).

Studies of growth and development in the area of human intelligence are also relevant. Many of these studies have found evidence that Henry and Hulin (1987) argued supports an interpretation of generally decreasing predictive validities (Anderson, 1939; Humphreys & Davey, 1984). Ackerman (1989) has challenged the conclusions of Henry and Hulin (1987) and the previous conclusions of Alvares and Hulin (1972, 1973) about the ubiquity of decreasing predictive validities.

The purpose of this analysis and article is to determine if predictive validities in general vary systematically as a function of time, stage of practice, or length of time on a job. Specifically, we are concerned whether temporally more remote performance assessments may be less strongly

related than temporally close performance to abilities assessed before performance or training and used as predictors of future performance.

Validity Generalization

We shall not review general validity evidence provided by primary empirical studies and subsequent meta-analyses of these studies. Meta-analyses of empirical studies of predictive validities of ability measures predicting performance have been carried out by, Schmidt and Hunter (1977), and their colleagues, (Pearlman, Schmidt, & Hunter, 1980; Schmidt, Gast-Rosenberg, & Hunter, 1980; Schmidt & Hunter, 1977). We do present, however, a brief synopsis of the past research in this area to establish a framework for our analyses and interpretations.

Individuals

The primary conclusion from past work on validity generalization is that validity estimates generalize across sub-populations of individuals, abilities/tasks, and situations. Schmidt and Hunter (1981) claim: "Professionally developed cognitive ability tests are valid predictors of performance on the job and in training for all settings" (1981, p. 1128). Analyses by Drasgow (1982) and Drasgow and Kang (1984), however, have raised questions about the power of most analyses of differential validity across sub-populations of individuals.

Tests and Tasks

Investigations of differential validity, the variance in validities of a given ability measure for different criterion tasks, have examined correlations between many combinations of ability tests and criterion task performance. Although there are disagreements about the appropriate conclusion, the results indicate that there are small but statistically reliable differential validities for some tests and task combinations. Measures of clerical/scholastic ability correlate more strongly with performance measures from a job family composed of clerical tasks than they do with performance measures in a mechanical job family (Humphreys, 1979). Conversely, ability measures based on mechanical/practical tests (Humphreys, 1979; Thurstone, 1938; Vernon, 1950) correlate more strongly with performance measures on mechanical tasks than they do with performance in clerical jobs (Humphreys, 1979).

Aside from this small but reliable difference in predictive validities between certain test-task combinations, there is little evidence for differential validity within broad job families. The observed differential validities are theoretically important. They may be, however, of limited practical utility. A test of general cognitive or intellectual ability will usually have a significant predictive validity for early performance on many jobs (Hunter, Schmidt, & Jackson, 1982; Schmidt, Gast-Rosenberg, & Hunter, 1980; Schmidt & Hunter, 1977, 1981; Schmidt, Hunter, & Caplan, 1981).

Situations

Situations, the final facet normally considered in validity generalization studies, have typically been investigated by studying validity across organizations as elements of a population of situations. The assumption is that small differences in situational variables, often instantiated as organizational climate (Schneider & Bartlett, 1968), would moderate test validity. Meta-analyses have demonstrated that the variance in observed empirical predictive validities across situations often can be accounted for by three sources of artifactual variance; sampling variance, unreliabilities, and restriction of range (Hunter, Schmidt, & Jackson, 1982; Schmidt & Hunter, 1977, 1981). After correcting for these three artifacts, there is little variance in empirical validities left to be explained by systematic differences among the elements of the populations of settings or situations.

In summary, there is some evidence for small but statistically reliable differential validities for some tests and job families. There is little evidence for variance in empirical validities across subpopulations of individuals, although the power of most analyses to detect substantial amounts of measurement bias is very low. There is also little evidence for systematic variance of validities across situations or organizations.

Generalization Across Time

Time has seldom been explored as a source of systematic variance in test validities. Variance of predictive validities across the time facet is theoretically and practically important; it addresses important questions related both to the stability of individual differences in abilities as well as dynamic vs. static criterion measures (Austin, Humphreys, & Hulin, 1989; Barrett, Caldwell, & Alexander, 1985; Ghiselli, 1956). The stability of both abilities and performance has implications for the scientific study of human behavior that extends beyond the immediate, narrow question of predictive validity generalization across time. These implications will be addressed in the discussion section.

Theoretical Importance

Variance in predictive validity across time is theoretically important for the study of individual differences in human abilities. There are three possibilities that should be considered. Predictive validities may be constant, within the limits of sampling fluctuations, across time. Predictive validities may vary randomly beyond the limits of sampling fluctuations. Predictive validities may vary systematically across time showing significant linear or higher order temporal trends.

If predictive validities are constant, initial predictions determined from regression equations for early performance may be used for predictions of later performance and provide reasonable bases for forecasting very long term performance and ability. If predictive validities vary randomly across time, beyond the limits of sampling fluctuations, then there may be no linear temporal factor involved in the variability. Other factors such as

unreliability of performance, rapidly changing motivation, or social competition factors may be responsible for the observed fluctuations. If, as the third alternative suggests, predictive validities vary systematically and not randomly across time, then human abilities may also change systematically. Changes in rank orders of individuals would be more likely than stability along any given ability dimension. Such a dynamic conception of human ability is not new (Alvares & Hulin, 1972, 1973; Dunham, 1974; Humphreys, 1968). This dynamic interpretation of human abilities has further implications for the definition of human ability and for distinctions between human abilities on the one hand and skills and knowledge on the other (Ackerman, 1989; Henry & Hulin, 1987, 1989). This is a fundamental issue in this area; definitions and implicit assumptions about human abilities determine many of the conclusions about the data reviewed in this article.

Validities that change systematically across time perhaps should lead us to question assumptions we make about intellectual, cognitive, or psychomotor abilities defined as fixed capacities. Whatever the theoretical basis of assumptions about fixed capacities--genetic determinants or events during critical periods of development--the assumptions and the theories may need revising. This fundamental assumption of human abilities needs to be made explicit and its implications examined empirically whenever possible. Humphreys (1985) and others (e.g., Wesman, 1956) have suggested that abilities are neither fixed nor are they capacities; to define them in that manner makes little sense theoretically or psychometrically.

Dynamic criteria represent the other side of the function linking individual differences in abilities to individual differences in performance. Just as we often make assumptions about the stability and generality of human abilities, we make parallel assumptions about the stability and specificity of skilled performance (Rothe, 1946a, 1946b, 1947, 1951, 1970, 1978; Rothe & Nye, 1958, 1959). These assumptions may also need to be reexamined. That is, rank orders of individuals in terms of skilled performance, even after group means and variances have stabilized, may be less constant than is commonly assumed.

If rank orders of individuals in terms of levels of skilled performance change systematically, a conceptualization of criterion performance in which the amounts of the abilities required for performance on the criterion task change systematically as a function of practice on the task is also possible. This second view of skilled performance has been offered previously as an explanation for changing decreasing predictive validities (cf. Fleishman & Hempel, 1954).

Practical Importance

Temporal trends in predictive validities are also of practical importance. Estimates of the utility of testing and selection programs are often based on extrapolations from the validities of tests for predicting performance during training, or early in an individual's working career. If predictive validities vary systematically across time, then extrapolating utility estimates beyond the initial observation periods may lead to serious

errors. Periodic retesting of individuals' abili y levels to update the information in prediction equations and to generate new predictions of performance on the basis of periodic ability assessments may need to become a standard part of personnel selection programs. As an alternative, we may need to recognize that our ability to predict long term performance is very limited; more modest claims for utility or predictive validities may be needed.

In summary, a more thorough understanding of predictive validity should include investigations of change across time and practice on the task, as well as differences across subpopulations of individuals, abilities, tasks/jobs, and situations. This relative lack of emphasis on the time facet should be rectified if we are to develop dynamic models of ability-performance relationships.

Goals of the Study

The goals of this study are to examine temporal trends in predictive validity coefficients within studies and to accumulate estimates of these temporal trends across studies. This is done by examining trends in the validities of tests for predicting individual differences in criteria at different stages of practice or performance within each study. After examining the temporal validity trends within each study and correcting for relevant statistical artifacts, the results are combined across studies. General and consistent temporal trends in predictive validities across studies would suggest restrictions on the generalizability of predictive validities. Evidence of long term as well as short term validity of tests as predictors of performance is needed for complete statements about validity and utility.

A second category of studies was included in addition to the set of studies reporting predictive validities as normally defined. These studies investigated the stability of ability or performance measures across time. The relevant data from such growth and development studies are usually presented as a time period-by-time period or trial-by-trial matrix of performance intercorrelations. The elements of the vector defined by the first row of such a matrix represent the validity of performance on the first trial, or during the first time period, for predicting performance during the 2nd, 3rd, and subsequent $n - 1$ trials or time periods of the task. As such, it is analogous to a validity sequence extracted from the usual predictive validity studies using ability measures to predict performance during sequential trials or in sequential time periods. We do not claim that performance during the first trial or first time period on a skilled task or an ability assessment is identical to the usual ability measures used as predictors of skilled performance. We do argue that distinctions between first trial performance measures and a job sample taken before hiring and used as a predictor of job performance are more apparent than real. We maintained the distinction between the typical predictive validity studies and the growth and development studies by analyzing and reporting the results separately.

In one important aspect, the analysis reported in this article is substantially different from standard meta-analyses. Because the time facet is ordered and linear, at least within the limits of the studies reviewed, the validity coefficients obtained in any study can be ordered along this dimension; time provides both the facet and the metric for ordering the observations for trend analyses. This characteristic of time enables us to go beyond simply estimating the variance in validity coefficients due to time or practice on the task. We can order the obtained predictive validity coefficients and test for systematic temporal trends. There is no compelling reason to assume only linear temporal trends in predictive validities, but there are normally not enough observations within any one study to estimate any higher order trends. Therefore, our study is limited to the examination of simple linear trends.

Artifact Corrections

Corrections will be made for the subset of the possible artifactual influences that can affect observed trends in predictive validities within each study. After making these corrections, the within study temporal trends will be accumulated across studies.

If there were differential reliabilities of performance measures across the time intervals or observations within a study, we corrected the observed validity coefficients for differential attenuation. This was necessary in order to estimate the temporal trend in predictive validity coefficients within each study unconfounded by systematic trends that might exist in performance reliability.

We also corrected the observed validity coefficients for differential range restriction across performance assessments within studies. Differential range restriction across observations, specifically decreases in variance across observations, has been suggested as an explanation for observed decreasing validities across time (Barrett et al , 1985). Correcting for differential range restriction will allow an investigation of this hypothesis.

This use of correction for range restrictions is somewhat different than the usual use of such corrections. Normally, corrections for range restrictions are for the purposes of estimating population validities from sample validities where the sample may be more or less variable than the population to which one wants to generalize. Differential range restriction across samples can introduce artifactual variance in sample validity coefficients (Hunter et al., 1982). Such artifactual variance must be removed in secondary analyses testing hypotheses about situational variance in validity.

In this study, we are not concerned about generalizations to populations of individuals; those meta-analytic studies have been conducted. We are concerned about artifactual influences on variance across performance assessments. If there are ceiling effects on performance that become more restrictive as the sample of individuals acquires more skill across the different assessments in a study, then the variance in

performance will be artifactually restricted across time. Predictive validities will appear to be lower for later assessments. Floor effects during early performance that become less serious as practice or performance continues may lead to increasing variance and artifactually increasing validities. Misleading conclusions may be reached about differential validity when the appropriate conclusions should be about differential range restrictions across trials or performance assessments. We corrected validity coefficients within studies to reflect differential range restrictions reflected by differential variance of performance across time. Our intent was to obtain estimates of predictive trends within studies that were not influenced by such artifacts. The details of the corrections for range restriction and unreliability, as well as our methods for obtaining the studies in our sample are described in the method section below.

II. METHOD

Our search of the literature on ability-performance relations spanned the areas of prediction of performance as well as growth and development research. Included in the performance prediction domain were experimental studies as well as studies of academic performance. The growth and development research included any longitudinal investigations of ability/performance in which intertrial correlations were reported. Many of these were studies of intellectual abilities. Overall 41 articles were collected yielding 77 independent validity sequences.

Data Collection Procedures

The collection of relevant empirical studies began with a search for review articles in the various subareas mentioned previously. Those used included Ackerman (1987), Adams (1987), Alvares and Hulin (1972), Barrett, Caldwell, and Alexander (1985), Guion and Gibson (1988), and Henry and Hulin, (1987). Because the scope and focus of these articles varied, many empirical studies cited in these articles were not relevant to our investigation, and some that appeared relevant according to the reviewing authors' descriptions did not include the necessary information for use in our analyses. Several articles not included contained relevant data but presented results only in the form of a graph of predictive validity or stability coefficients against time or the trial's ordinal position (e.g., Stelmach, 1969). To be included we would have had to estimate the validity or stability coefficients by extrapolating from the graph. Because most of these studies were cummulative, rather than uniquely informative, they were not included.

Potential studies were examined to assess their appropriateness for secondary analyses. Two conditions were necessary for inclusion: (a) longitudinal correlational analyses (not cross-sectional), and (b) at least three correlations between ability and performance at different times representing predictive validity coefficients. Cross-sectional studies were not included because investigations of predictive validities across time necessitate the use of longitudinal designs. Longitudinal designs are required because of the nature of some of the hypotheses concerning changing

validities (e.g., that rank orders of individuals change in terms of abilities). If cross-sectional samples are used, the effects of changing rank orders of individuals cannot be investigated.

Studies were not included if only two validities were reported (e.g. Adams, 1957). Others were omitted because only factor loadings of performance on extracted dimensions were given. Of the studies included, some reported multiple sequences of validities using a different predictor for each sequence. These were included as separate validity sequences in our analysis. However, if multiple sequences of validities were based on subsets of the total sample, say, males and females, only the total sample validity sequence was used (if reported).

Although the search was systematic and extensive, there are undoubtedly studies that were not located. If these unlocated and unanalyzed studies are systematically different in terms of temporal trends in validity then the conclusions reported here may be more general than the data warrant. It is unlikely, however, that our study is plagued by "file drawer" problems (Rosenthal, 1979); there should be no systematic effect on the publication of studies in which predictive validities are stable, increase or decrease systematically, or vary widely but randomly. Temporal trends in validity coefficients have rarely been the main topic of interest in most studies of predictive validity. Positive, negative, or even zero trends in predictive validities, by themselves, should not directly influence decisions by investigators to submit manuscripts for publication or by editors to accept or reject these manuscripts.

Statistical Analyses

The first phase of the analyses consisted of plotting the observed predictive validities against time. For this analysis, time was treated simply as an ordinal variable. Regression lines were fitted and the slopes calculated for the within study regression of predictive validity or stability coefficients on time. For evidence of non-linearity, all such plots were examined visually since most studies did not include a sufficient number of data points to permit statistical analyses. There was little non-linearity evident in these plots. In addition, an index of validity change across observations within studies was calculated by computing the difference between the two endpoints of the regression line. These are the predicted validities that corresponded to the first and last observations in the sequence. This difference represents the amount of decrease (negative Δr) or increase (positive Δr) in validity as a function of time and practice on the task. We used the difference between the predicted validities corresponding to the first and last points on the regression line, Δr, rather than the raw difference between the first and last coefficients, Δr, to remove as much as possible the effects of within study fluctuations in the validity or stability sequence caused by sampling variance.

In the second phase, each observed predictive validity estimate was corrected for range restriction, unreliability, and if possible, for both. Changes in standard deviations of performance across the assessments in each study were used to correct for differential range restriction.

Standard deviations or variances were reported in 35% of the studies. If not reported, no correction was made. Corrections for differential range restriction were made by first calculating an average standard deviation across trials, weighted by sample size. This weighted average was then used in the formula for correcting for range restriction.

These corrected validities were also regressed on time and the changes in the corrected predictive validities corresponding to the first and last observations calculated (Δr^*; * indicates that the predictive validities have been corrected for any of the statistical artifacts that were possible to correct for given the available data). Differences in the changes in the regressed uncorrected and the corrected validities onto time reflect the effects of statistical artifacts on changes in the validities across time. The changes in the corrected validities, Δr^*, where available, were used in all subsequent analyses. If reliabilities were not included (as was the case 90% of the time) correlations between adjacent trials were used to estimate reliability in the correction formula. The square roots of these correlations were used in the denominator of the Spearman-Brown formula. If corrections for range restriction had been made in the previous step, these corrected correlations were used in the numerator of the Spearman-Brown formula to correct for unreliability. If the correction for range restriction could not be made, the uncorrected correlations were used. If neither reliabilities nor adjacent trial correlations were included, a correction for unreliability was not made.

A final index of change in validity sequences as a function of time was computed by correlating validity (corrected for unreliability and range restriction, if provided) with the ordinal time variable within each study.

The third phase of the analysis consisted of combining within study temporal trends in validity coefficients across studies to provide an overall estimate of the trends in predictive validities. This analysis was not straightforward. Choosing a reasonable metric to represent time that was both sensitive to small time differences within studies and also made sense across studies is difficult; the time periods ranged from indices of scientific productivity across decades to several 1 or 2 minute measures of performance on psychomotor tasks across a 1-hour experiment. We used two different representations of time: a natural log transformation of time and a simple ordinal time-metric. The former may assign unrealistically large values to later observations in a very long term longitudinal study; the latter discards information by providing only rank order values and may assign unrealistically small values to later observations in long-term longitudinal studies. We report Δr^* regressed on both rank-ordered time and the natural log of time.

III. RESULTS

Table 1 presents the summaries of our predictive validity results. The columns from left to right represent: (a) the number of observations or data points for each validity sequence, (b) the amount of time elapsed during the period of data collection, (c) N = the number of subjects, (d) Δr^a = the decrement in validity corrected for range restriction, (e) Δr^b =

the decrement in validity corrected for unreliability, (f) Δr^c = the decrement in validity corrected for both sources of artifactual variance, (g) the correlation between time or assessment period and validity, (h) time elapsed in hours, and (i) the natural log transformation of time elapsed.

Of the prediction studies, 82% (44 of 54) showed decreasing validity patterns as measured by the change in validity (uncorrected) as calculated from the regression of validity against time (Δr). When the validities were corrected for range restriction (Δr^a), the decrease in Δr became even more pronounced 47% of the time (7 of 15). Similarly, when the observed validities were corrected for unreliability (Δr^b), the decrease in Δr was stronger 81% of the time (25 of 31). Correcting for both statistical artifacts (Δr^c) yielded a more negative Δr in 86% of the cases in which it was possible to correct for both artifacts (12 of 14). Overall, only 10 validity sequences yielded positive or zero uncorrected Δr's; none of the validity sequences yielded positive or zero Δr's when corrections were made for both unreliability and range restriction. The average corrected and uncorrected Δr are given at the bottom of Table 1. The average correlation between corrected predictive validities making up the validity sequence and the temporal rank order of the observation was -.80 (r to z transformation weighted by the number of observations). This average correlation represents the degree of within study correlation between the ordinal position of the observation within the study and the predictive validity of the test being used to forecast task performance. Both the size of this correlation and a perusal of Table 1 suggest a great deal of consistency in the relation between temporal position of performance and the predictive validity of tests across a variety of tasks, populations, and situation.

The average decrements in the validity coefficients range from -.15 when no corrections were made to -.60 when corrections could be made for both differential range restrictions and attenuation within studies. The 90% confidence intervals for those decrements that could be corrected for at least one of the potential statistical artifacts never included zero; none of the individual values of the decrements in corrected validity coefficients were zero or positive. The value calculated for the average corrected validity decrement, -.60, may not represent the best measure of central tendency of distribution of corrected validity decrements because of one extreme value, -1.21. The median of the distrubution of corrected validity decrements is -.29; the mean of the distribution after discarding the two most extreme values, -1.21 and -.10, is -.45. Either of these latter estimates of central tendency, although somewhat discrepant, probably represents a more accurate summary measure of the central tendency of the distribution. The mean of the trimmed distribution, -.45, is more consistent with the overall information contained in this analysis.

The average within study correlation (z-transformation, weighted by the number of data points within the study) between the time of the performance assessment and the validity of the test for predicting that performance assessments was -.80. This correlation is highly significant and attests to the consistency and significance of the validity decrement across time within each study.

Table 1. Summary of Predictive Validity Results

AUTHORS, DATE Predictor	# Observations	Time elapsed	N	Δr uncorrected	Δr^a	Δr^b	Δr^c	Correlation between time and r	Time elapsed (hours)	ℓn (time)
ALVARES & HULIN, 1973										
Pretest, flight training	3	15 wks.	61	-.23	---	---	---	-1.00	2,520	7.83
Postest, flight training[d]	3	15 wks.	67	-.05	---	---	---	- .75	2,520	7.83
BILODEAU, 1952										
Two-handed coordination test	7	12 min.	152	-.09	-.16	-.16	- .23	- .93	.20	-1.61
BILODEAU, 1953										
Micrometer reading	15	15 min.	40	-.33	-.18	---	---	- .43	.25	-1.39
BILODEAU & RYAN, 1960										
Line drawing	15	30 min.	48	-.26	-.08	-.24	- .10	- .21	.50	- .69
BRENNER & LOCKWOOD, 1965										
Salary, engineers	22	21.5 yrs.	52	-.55	-1.03	-.73	-1.21	- .98	188,340	2.15
DUNHAM, 1974										
Visual Acuity[d]	5	1 hr.	48	-.01	---	-.01	---	- .87	1	0
Reaction Time	5	1 hr.	48	-.22	---	-.22	---	- .87	1	0
Tracking	11	1 hr.	48	-.25	---	-.26	---	- .73	1	0
FLEISHMAN, 1953										
6-Target Rudder Control	3	8 min.	342	-.14	---	-.17	---	- .99	.13	-2.01
6-Target Rudder Control (predicting Std. Rudder Control)	6	8 min.	342	.04	---	---	---	.22	.13	-2.01
Standard Rudder Control	5	8 min.	356	-.07	---	-.12	---	- .78	.13	-2.01
Standard Rudder Control (6-Target Rudder Control)	4	8 min.	356	-.14	---	---	---	- .96	.13	-2.01

Table 1. (Continued)

AUTHORS, DATE Predictor	# Observations	Time elapsed	N	Δr uncorrected	Δr^a	Δr^b	Δr^c	Correlation between time and r	Time elapsed (hours)	ℓn (time)
FLEISHMAN, 1960										
Rotary Pursuit Test	7	15 min.	224	-.13	---	-.28	---	- .96	.25	-1.39
FLEISHMAN & FRUCHTER 1960										
Learning Morse Code	3	66 days	310	-.01	-.29	-.31	-.59	- .93	1584	7.37
FLEISHMAN & HEMPEL, 1954										
Complex Coordination	7	2 days	197	-.17	---	-.23	---	- .88	48	3.87
FLEISHMAN & HEMPEL, 1955										
Discrimination Reaction Time	7	30 min.	264	-.10	---	-.22	---	- .97	.50	- .69
FLEISHMAN & RICH, 1963										
Aerial Orientation	10	1 hr.	40	-.38	---	---	---	- .94	1	0
Kinesthetic Sensitivity[d]	10	1 hr.	40	.30	---	---	---	.83	1	0
HENRY & HULIN, 1987										
Annual Total Runs Produced	9	10 yrs.	94	-.10	---	-.21	---	- .54	87,600	11.38
Overall Pitcher Performance	9	10 yrs.	38	-.30	---	-.44	---	- .75	87,600	11.38
HINRICHS, 1970										
Pretest (Speed/Speed)	3	1 hr.	27	.12	---	---	---	.87	1	0
Pretest (Accuracy/Speed)	3	1 hr.	27	-.32	---	---	---	- .96	1	0
Pretest (Speed/Accuracy)	3	1 hr.	23	.30	---	---	---	.97	1	0
Pretest (Accuracy/Accuracy)	3	1 hr.	23	-.55	---	---	---	- .98	1	0
HOLLANDER, 1965										
Peer Nomination	3	12 wks.	639	-.19	---	-.28	---	- .98	2,016	7.61

Table 1. (Continued)

AUTHORS, DATE Predictor	# Observations	Time elapsed	N	Δr uncorrected	Δr^a	Δr^b	Δr^c	Correlation between time and r	Time elapsed (hours)	ℓn (time)
HUMPHREYS, 1960										
Grade Point Average (GPA)	7	4 yrs.	91	-.33	---	-.41	---	- .96	35,040	10.46
HUMPHREYS, 1968										
GPA	7	4 yrs.	1610-7255	-.27	-.21	-.30	-.24	- .97	35,040	10.46
GPA	7	4 yrs.	1600	-.17	-.34	-.24	-.27	- .97	35,040	10.46
HUMPHREYS & TABER, 1973										
GPA	7	4 yrs.	1549-3015	-.26	-.24	-.31	-.29	- .92	35,040	10.46
JONES, 1970										
2-hand coordination test	5	6 days	38	-.09	-.23	-.14	-.28	- .98	144	4.97
JONES, KENNEDY & BITTNER, 1981										
Video game scores	5	6 days	22	-.02	---	-.09	---	- .38	144	4.97
KAUFMAN, 1972										
Ability Predicting Performance	3	9 yrs.	110	.04	---	---	---	.99	78,840	11.28
Ability Predicting Papers	3	9 yrs.	110	.02	---	---	---	.62	78,840	11.28
Ability Predicting Patents	3	9 yrs.	110	.02	---	---	---	.56	78,840	11.28
LIN & HUMPHREYS, 1977										
GPA (physics, chem, math)	9	7 yrs.	250-1047	-.50	-.46	-.56	-.52	- .94	61,320	11.02
GPA (law)	9	7 yrs.	703-1230	-.41	-.37	-.53	-.49	- .98	61,320	11.02
GPA (law)	9	7 yrs.	674	-.32	-.34	-.40	-.50	- .93	61,320	11.02

Table 1. (Continued)

AUTHORS, DATE Predictor	# Observations	Time elapsed	N	Δr uncorrected	Δr[a]	Δr[b]	Δr[c]	Correlation between time and r	Time elapsed (hours)	ℓn (time)
NOBLE, 1970										
Rotary Pursuit Task	19	35 min.	500	-.33	---	-.35	---	- .95	.58	- .54
PARKER & FLEISHMAN, 1960										
Aerial Orientation	3	5 wks.	203	-.02	---	*-.03	---	- .63	840	6.73
Discrimination Reaction Time	3	5 wks.	203	-.02	---	*-.03	---	- .98	840	6.73
Complex Coordinator	3	5 wks.	203	-.01	---	*-.02	---	- .76	840	6.73
PAULSEN. 1935 (I)										
Steadiness Test	9	10 min.	97	-.13	-.13	-.02	-.15	- .80	.17	-1.79
PAULSEN, 1935 (II)										
Steadiness Test	8	18 days	80	-.18	-.03	-.02	-.26	- .26	432	6.07
POWERS, 1982										
UGPA, 23 Law Schools	3	3 yrs.	12,755	.00	---	---	---	.20	26,280	10.18
POWERS, 1982										
LSAT, 23 Law Schools	3	3 yrs.	12,755	-.08	---	---	---	- .89	26,280	10.18
REYNOLDS, 1952										
Complex Coordination	11	8 hrs.	153	-.22	-.35	-.03	-.38	- .89	8	2.08
ROTHE, 1946										
Butter Wrapper Work Curves	4	7 days	8	-.14	---	---	---	- .50	168	5.12
VITELES, 1983										
Serial Discrimter	4	3 days	84	0	---	---	---	- .09	72	---
Brown Spool Packer	4	3 days	84	-.32	---	---	---	- .98	72	---

Table 1. (Concluded)

AUTHORS, DATE Predictor	# Observations	Time elapsed	N	Δr uncorrected	Δr^a	Δr^b	Δr^c	Correlation between time and r	Time elapsed (hours)	ℓn (time)
WINTERBOTTOM, PITCHER, AND MILLER, 1963										
Pre-Law Record, School A	3	4 yrs.	74	.03	---	---	---	.98	35,040	10.46
LSAT, School A	3	4 yrs.	74	-.04	---	---	---	- .24	35,040	10.46
Pre-Law Record, School B	3	4 yrs.	74	.02	---	---	---	.28	35,040	10.46
LSAT, School B	3	4 yrs.	74	-.23	---	---	---	- .83	35,040	10.46
Average				-.15	-.52	-.25	-.60	- .80		
# of Negative Entries				44	15	31	14	44		
# of Positive Entries				9	0	0	0	10		
# of Zero Entries				1	0	0	0	0		
Total Number of Entries				54	15	31	14	54		

[a]corrected for differential range restriction.

[b]corrected for unreliability.

[c]corrected for both differential range restriction and unreliability.

[d]abilities hypothesized to be stable by the primary investigators.

Table 2 shows the summaries of the intertrial performance studies in which the same performance measure was assessed across trials or time periods. These studies are fewer in number, but the results provide even stronger evidence for the trends found in Table 1. All of the Δr's, both corrected and uncorrected are negative. The correlations between time and r are also in every case negative. As with the prediction studies, correcting for unreliability or range restriction, or both had the effect of making the decrease in Δr even greater in 22 of the 23 cases. The average correlation between corrected predictive validities and time was -.94 (r to z transformation, weighted by number of within study data points).

Table 2. Summary of Stability of Performance Results

AUTHORS, DATE Variable/Performance	# Observations	Time elapsed	N	Δr uncorrected	Δr^a	Δr^b	Δr^c	Correlation between time and r	Time elapsed (hours)	£n (time)
ANDERSON, 1939										
IQ Test Score	5	5 yrs.	150	-.05	-.26	---	---	- .95	43,800	10.69
Mental Age--Boys	9	9 yrs.	135	-.13	-.29	---	---	- .90	78,840	11.28
Mental Age--Girls	9	9 yrs.	130	-.10	-.29	---	---	- .84	78,840	11.28
Butler & McCauley, 1987										
GPA	3	4 yrs.	618	-.06	-.05	-.02	-.07	-1.00	35,040	10.46
GPA	3	4 yrs.	631	-.04	-.04	-.02	-.06	-1.00	35,040	10.46
DENNIS, 1954										
Scientific Productivity, Psychologists	4	50 yrs.	43	-.01	---	-.01	---	- .01	438,000	12.99
Scientific Productivity, Scientists	4	50 yrs.	41	-.46	---	-.63	---	- .99	438,000	12.99
DENNIS, 1956										
Scientific Productivity, Octogenarians	5	60 yrs.	56	-.24	---	-.40	---	- .99	613,200	13.33
GOUGH & HALL, 1975										
GPA, Med. School	3	4 yrs.	601	-.34	-.35	-.09	-.44	-1.00	35,040	10.46
HENRY & HULIN, 1987										
Annual Total Runs (Majors)	8	9 yrs.	94	-.27	---	-.37	---	- .70	78,840	11.28
Pitcher Performance (Majors)	8	9 yrs.	38	-.18	---	-.30	---	- .69	78,840	11.28
HUMPHREYS, 1960										
Shorthand Sessions	7	8 days	28	-.20	---	-.37	---	- .84	192	5.26

Table 2. (Continued)

AUTHORS, DATE Variable/Performance	# Observations	Time elapsed	N	Δr uncorrected	Δr[a]	Δr[b]	Δr[c]	Correlation between time and r	Time elapsed (hours)	ℓn (time)
HUMPHREYS & DAVEY, 1984										
Mental Test Scores	11	15 yrs.	49-590	-.39	---	-.48	---	- .93	131,400	11.79
Intelligence Test Scores	11	11 yrs.	273-1204	-.18	---	-.24	---	- .94	96,360	11.48
HUMPHREYS & PARSONS, 1979										
Listening	3	6 yrs.	1430	-.06	---	-.10	---	- .85	52,560	10.87
Composite	3	6 yrs.	1430	-.03	---	-.07	---	- .98	52,560	10.87
PARKER & FLEISHMAN, 1960										
Integral Error Score	9	6 hrs.	203	-.44	-.44	-.16	-.60	- .81	6	1.79
Horizontal Error Score	9	6 hrs.	203	-.40	-.40	-.19	-.59	- .88	6	1.79
Vertical Error Score	9	6 hrs.	203	-.41	-.40	-.17	-.57	- .85	6	1.79
Sideslip Error Score	9	6 hrs.	203	-.45	-.45	-.17	.62	- .90	6	1.79
Time on Target Score	9	6 hrs.	203	-.51	-.49	-.19	-.68	- .98	6	1.79
ROFF, 1941										
Stanford-Binet Scores	3	4 yrs.	45	-.02	-.15	-.05	-.18	-1.00	35,040	10.46
WILSON, 1983										
Mental Test Scores	13	15 yrs.	988	-.32	-.24	-.48	-.40	- .74	131,400	11.79

Tab . (Concluded)

	Δr uncorrected	Δr[a]	Δr[b]	Δr[c]	correlation between time and r
Average	-.24	-.30	-.24	-.45	-.94
# of Negative Entries	23	13	20	10	23
# of Positive Entries	0	0	0	0	0
Total # of Entries	23	13	20	10	23

[a]corrected for differential range restriction.

[b]corrected for unreliability.

[c]corrected for both differential range restriction and unreliability.

IV. DISCUSSION

Previous secondary analyses have investigated the generalizability of validities across populations, situations, abilities, and tasks. These analyses have concluded that observed variance in test validities across these populations is substantially due to statistical artifacts. Some researchers have been willing to argue that validities generalize across these facets almost without limit (Schmidt & Hunter, 1977, 1981). In contrast, our secondary analysis of predictive validities across time has demonstrated that validities should not be generalized across this facet of validity. Time, a relatively unstudied facet in validity generalization research, has a consistent effect on predictive validities. Validities vary across time; with few exceptions, they decrease monotonically.

This secondary analysis began with a specific set of hypotheses about the nature of variance in predictive validities. We were not concerned with simply estimating variance due to artifacts or design features of the studies in our sample. We were able to formulate specific hypotheses on the basis of previous summaries of the literature (Alvares & Hulin, 1972, 1973; Henry & Hulin, 1987). These hypotheses addressed the nature of the variance of predictive validities; predictive validities should decrease monotonically with time. Failure to reject the null form of this hypothesis (i.e., no temporal, monotonic decrease) is more informative than rejecting a simpler hypothesis that the variance in predictive validities is greater than would be predicted by sampling fluctuations, differences in reliabilities, and differences in variance of performance assessments.

Past researchers who have discussed decreasing predictive validities across time in organizational settings have attributed the observed decrement to statistical artifacts (Barrett et al., 1985). That is, differential range restriction and unreliability across different time periods or trials were the putative reasons for the observed decreases (Barrett et al., 1985). Among the studies providing information necessary to correct for either or both of these statistical artifacts, 84% (27/32) of the prediction studies shown in Table 1 and 96% (22/23) of the stability, growth, and development studies shown in Table 2 revealed that predictive validities decreased more when corrected for these artifacts than did the uncorrected validities. None of the studies that have positive slopes of the regression of predictive validities onto time contained information necessary to correct for the artifacts; it is unknown if these slopes would remain positive if the validities could be corrected for differential unreliability and range restriction across performance assessments.

This finding of greater temporal variance following corrections for statistical artifacts stands in sharp contrast to other secondary analyses of variance in predictive validities across facets or populations. These previous analyses have found observed variance in predictive validities was generally attributable to artifacts. Our analyses revealed that removal of the statistical artifacts increased the negative slopes of the regressions of validity onto time and, hence, the variance of predictive validities accounted for by time.

The pervasiveness of this systematic decrease in validities can be seen by reviewing Tables 1 and 2. Forty-four of the 54 validity sequences included in Table 1 and 23 of the 23 validity sequences in Table 2 had negative slopes for the regressions of predictive validity onto time. The average within study correlations between predictive validity and the ordinal position of the performance assessment was -.80 in Table 1 and -.94 in Table 2. The number of observations in the studies ranges from 3 to 22. The durations of the studies are from 8 minutes to nearly 22 years among the prediction studies, and as long as 60 years among the stability, growth, and development studies. The types of abilities investigated range from specific and narrow (e.g., simple reaction time, discriminant reaction time) to broad and general abilities (e.g., general intellectual ability). The performance predicted in the studies ranged from the specific (Pursuit rotor performance) to the very general (flight performance). Populations sampled covered highly selected groups in terms of abilities and skills being studied (professional baseball players) to samples from student populations. Laboratory and field studies were both well represented. There were few exceptions to the observed decreasing trends in predictive validities.

The one striking exception to the trends observed in the data in Table 1 is found in a series of studies conducted by Powers (1982) and Winterbottom et al., (1963) predicting grades in law schools using undergraduate grades and Law School Aptitude Test scores as the predictors. These studies found that although LSAT validities declined consistently across the 3 years of law school, the validity estimates for the undergraduate grades did not show the expected validity decrement. Both of these trends, the negative temporal trend in LSAT validities and the zero or slightly positive trends in the predictive validity of undergraduate grades, were consistent across more than 20 different law schools. The difference in the validity sequence trends suggests the zero or positive slope for the validity of undergraduate grades cannot be attributed to criterion contamination or related criterion problems. The same criterion resulted in opposite trends in the same sample of law schools.

There is no obvious explanation for the discrepant trends found in law school grades nor is there any obvious explanation for the difference in the trends between LSAT scores and undergraduate grades as predictors. In spite of an apparent finding of generality across situations reported by Schmidt and Hunter (1977, 1981), law schools may represent a significantly different situation for temporal generalizations.

The regression of Δr on the number of observations per validity sequence showed that across studies, decrements in validity became more pronounced as the number of data points increased. Across the prediction studies in Table 1, this regression was -.51; across the studies in Table 2, this regression was -.38.

Other deviations from the overall trends in Table 1 should also be noted. Fleishman and Rich (1963) reported an increasing correlation between kinesthetic sensitivity and psychomotor performance. This increasing, as opposed to a decreasing, correlation was predicted by these authors on the

basis of a conceptual explanation for the generally observed decreases in predictive validities that stressed changes in abilities required for performance as a function of practice on the task.

Hinrichs (1970) reported generally increasing correlations between pretest measures and performance across different trials on a psychomotor task. Although one of these increasing validity sequences had been predicted by Hinrichs, the extreme amount of within study fluctuation in predictive validities from trial to trial and the very small sample size make the significance of the trends difficult to interpret.

Three additional increasing validity sequences were reported by Kaufman (1972). These increasing validity sequences involved scientific performance measures including papers written and patent disclosures. Both criterion contamination and situational variance may partially account for these discrepant findings.

In general, aside from the undergraduate grade point average predicting law school grades and the increasing validity of a measure of kinesthetic sensitivity for predicting psychomotor performance, the discrepancies to the observed general trends in predictive validities seem to represent anomalies more than significant departures from general findings that need to be explained. Replication of the increasing validity sequence for psychomotor performance needs to be done. If the increasing trend is replicated, it should lend support to an explanation of changes in validities being caused by changes in abilities required for performance on the task.

We have not attempted to weight estimated "effect" sizes by sample sizes for each study to obtain an expected effect size. This weighting procedure is justified when the effect sizes being estimated have some meaning when applied to individuals. That is, if the effects represent the expected change that may occur in an individual as a result of the experimental manipulation or naturally occurring event, such weighting and estimation procedures are reasonable. The dependent variables analyzed in this secondary analysis were correlations and changes in correlations that have a meaning for a study or for a group as an undifferentiated whole; they have no meaning in this context when disaggregated to individual data.

Although we analyzed the effects of time on validity, we do not imply that time per se was the causal factor in the observed validity decrements. Those things that occur while individuals are learning and performing jobs and during skill acquisition are the assumed causal agents. Time is necessary to allow these things to occur and is a convenient metric in the absence of more specific indicators. Studies of the effects of the specific events indexed by time are the obvious next steps in this area.

The theoretical and practical implications of these findings need to be addressed in detail in laboratory and field studies. In this paper, given that our goals were to establish the existence and form of any temporal relationship with predictive validities, we can discuss them only briefly (for a more in-depth treatment, see Alvares & Hulin, 1973; Henry & Hulin, 1987).

Theoretical Implications

Two theoretical explanations for the observed decrement in predictive validities were discussed by Alvares and Hulin (1972, 1973). The explanatory power of the two explanations were compared in an experimental study of pursuit rotor performance by Dunham (1974). Briefly, Fleishman (1960) advanced an explanation for the observed decrement that stressed changes in the combination of abilities required to perform the task. These hypothesized changes in abilities required by the task occur as a result of practice and increasing task proficiency. Adams (1957), Alvares and Hulin (1972, 1973), and Bechtoldt (1960, 1961) have discussed flaws in the empirical support offered by Fleishman (1960) for this explanation of validity decrements.

An alternative explanation stressing changes in individuals' abilities, as a function of practice on tasks requiring those abilities, was discussed in detail by Alvares and Hulin (1972, 1973). This explanation explicitly rejects an assumption of fixed abilities. It assumes instead that individuals' ability levels change as a function of complex skill acquisition. Abilities have been defined as consisting of the current repertoire of relevant skills and knowledge possessed by an individual (Hulin & Humphreys, 1980; Humphreys, 1985; Wesman, 1956) rather than fixed capacities. Individuals' ability levels are assumed to undergo significant changes whenever they acquire proficiency in complex tasks. Further, the rank order of individuals in terms of their relevant skills and abilities does not remain constant during the process of skill acquisition. Some individuals exhibit greater changes than others in the abilities related to task performance. According to this explanation for validity decrements, the set of abilities required for task performance does not change; the amounts of the abilities individuals have change. Specifically, the amounts of the relevant abilities differ from early to late in performance or learning. Predictive validities of late performance that are based on early ability assessments (i.e., those taken before individuals begin a job or practice a task) are low because ability levels assessed before performance has started may be only moderately related to the ability levels individuals have late in performance.

A competitive test of these two explanations in terms of their ability to explain the decrement in predictive validities over time and practice showed that a number of the hypotheses based on the changing ability levels explanation were supported (Dunham, 1974). However, that explanation was not able to account for all of the observed validity decrement. A postdictive validity sequence consisting of the correlations between ability tests given after training on a task should have had a positive slope that mirrored the negative slope of the predictive validity sequence based on tests given before practice on the task. Although the postdictive validity for performance on the final trial was greater than the predictive validity for the final trial, it was not as high as the predictive validity for the initial trial. Similar results were obtained by Alvares and Hulin (1973). Dunham (1974) concluded that there was no empirical evidence supporting the

explanation based on changing task requirements; there was support for the explanation based on changing subject ability levels but it could not explain the entire validity decrement.

A third explanation emphasizes social factors in task performance and skill acquisition. That is, individual performance is hypothesized to be a function of two independent factors: relevant abilities and the ability or skill level of the group with which the individual is competing, learning the skill, or performing their job. This explanation assumes that individuals know the average performance level of the selected group of which they are a member. Those well below the average group performance on the task at any given time are expected to increase their efforts on the job or task; those well above the group average may slacken their efforts relative to other group members. Thus, regression to the mean of the selected group is offered as an explanation for validity decrements (L.G. Humphreys, personal communication).

This explanation has a great deal of appeal for explaining within group validity decrements that occur in groups that interact a great deal during training or on the job. Such selected groups as pilot trainees, law school students, and employees in an organization have this characteristic. Within group competition may be a powerful factor in influencing group members' performance levels. Other "groups" created in laboratory studies are little more than collections of individuals aggregated for purposes of data analyses. Interactions among the members of the experimental and control groups in most of these studies are minimal or nonexistent. The social competition explanation loses much of its intuitive appeal when applied to validity decrements observed in these experimental studies.

Practical Implications

The practical implications of these three theoretical explanations for observed validity decrements are substantially different. The first two (ability-based) explanations suggest that both predictive validities and the practical utility of selection programs decrease over time and are temporally limited to early performance on a task or job or to performance during training. If abilities required for late performance are independent of those required for early performance, and if, as our results suggest, nearly all commonly assessed abilities are those that are required for early rather than late performance, then both the within group predictive validity and utility of selection programs will decrease concomitantly. If abilities change significantly as a result of practice on the task, and if ability increments at time $i + 1$ are independent of ability level at time i, then after extensive practice on a task, ability levels should be nearly independent of ability levels used to select individuals. Extensive research by Humphreys (1960) and Humphreys and Davey (1984) has suggested that this hypothesized form of ability change cannot be rejected. Matrices of time period by time period correlations of ability levels generally show an excellent fit to a simplex matrix (Guttman, 1955).

If the third explanation, based on social competition among the members of the selected group, is correct, it suggests that decrements in predictive validities are not necessarily related to decrements in the utility of a selection test or program. As long as the regression is to the mean of the selected group, then the mean of the selected group may remain above the overall performance of the unselected population assuming the test was initially a valid predictor of performance in the overall population.

The changing ability level explanation offered for validity decrements suggests a need to develop theories of human ability and human performance that incorporate change. That is, rather than relying on static models of human ability in which ability levels are assumed to be fixed, dynamic models should be developed. These models would allow for systematic changes in ability level as a function of learning, or practice on, a complex skill. Those abilities required for task performance might be assumed to change as practice continues. Initial ability levels could be used to predict initial task performance. Performance on the task late in learning is assumed to be a function of the same abilities as those related to initial performance. However, either updated assessments of these abilities would be needed to predict later performance, or initial abilities plus a set of factors related to changes in abilities would be required to predict late performance. This set of factors related to change in ability would not necessarily be related to initial performance. Their relation to late performance would be through their effects on ability levels that were changed as a result of learning a complex task. The outcome of such a dynamic theory depends on identifying and assessing the set of individual or individual/environmental interaction factors related to ability change.

The changing task explanation for validity decrements requires a somewhat different strategy by researchers. Instead of searching for a set of factors that are related to ability change within individuals, this explanation would direct us to search for a set of abilities that are uniquely related to late performance on relevant criterion tasks. Regression equations predicting performance at different stages of practice would be characterized by a gradual decline in the sizes of the regression weights assigned to "early" abilities, those abilities related to initial performance levels, and a gradual increase in the sizes of regression weights assigned to "late" abilities. The outcome of this search for these "late" abilities is an empirical question. Past work by Fleishman (e.g., 1960), however, does not provide a great deal of encouragement for those interested in this line of inquiry. Ackerman (1987), however, has recently developed a theoretical framework consistent with this approach that may offer non-obvious insights and promise. However, given the variety of abilities studied as predictors of performance, any conceptual model based on unique human abilities that will predict late performance better than early performance faces long odds in its search for new abilities.

REFERENCES

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. Psychological Bulletin, 102, 3-27.

Ackerman, P. L. (1989). Within-task intercorrelations of skilled performance: Implications for predicting individual differences? (A comment of Henry & Hulin, 1987). Journal of Applied Psychology, 74, 360-364.

Adams, J. A. (1957). The relationship between certain measures of ability and acquisition of a psychomotor response. Journal of General Psychology, 57, 121-134.

Adams, J. A. (1987). Historical review and appraisal of research on the learning, retention, and transfer of human motor skills. Psychological Bulletin, 101, 41-74.

Alvares, K. M., & Hulin, C. L. (1972). Two explanations of temporal changes in ability-skill relationships: A literature review and theoretical analysis. Human Factors, 14, 295-308.

Alvares, K. M., & Hulin, C. L. (1973). An experimental evaluation of a temporal decay in the prediction of performance. Organizational Behavior and Human Performance, 9, 169-185.

Anderson, J. E. (1939). The limitations of infant and preschool tests in the measurement of intelligence. Journal of Psychology, 8, 351-379.

Austin, J. T., Humphreys, L. G., & Hulin, C. L. (1989). A critical reanalysis of Barrett, Caldwell, and Alexander. Journal of Applied Psychology, 42, 583-596.

Barrett, G. V., Caldwell, M. S., & Alexander, R. A. (1985). The concept of dynamic criteria: A critical reanalysis. Personnel Psychology, 38, 41-56.

Bechtoldt, H. P. (1960, April). Statistical tests of predictions generated from factor hypotheses. Paper presented at the Midwest Psychological Association, St. Louis, MO.

Bechtoldt, H. P. (1961). An empirical study of the factor analysis stability hypothesis. Psychometrica, 26, 405-432.

Bilodeau, E. A. (1952). Transfer of training between tasks differing in degree of physical restriction of imprecise responses. USAF Human Resources Research Center Research Bulletin, No. 52-54.

Bilodeau, E. A. (1953). Speed of acquiring a simple motor response as a function of the systematic transformation of knowledge of results. American Journal of Psychology, 66, 409-420.

Bilodeau, E. A. and Ryan, F. J. (1960). A test for interaction of delay of knowledge of results and two types of interpolated activity. Journal of Applied Psychology, 59, 414-420.

Brenner, M. H., & Lockwood, H. C. (1965). Salary as a predictor of salary. Journal of Applied Psychology, 49, 4, 295-298.

Butler, R. P. & McCauley, C. (1987). Extraordinary stability and ordinary predictability of academic success at the United States Military Academy. Journal of Educational Psychology, 79, 83-86.

Dennis, W. (1954). Predicting scientific productivity in later maturity from records of earlier decades. Journal of Gerontology, 9, 465-467.

Dennis, W. (1956). Age and productivity among scientists. Science, 123, 724-725.

Drasgow, F. (1982). Biased test items and differential validity. Psychological Bulletin, 92, 526-531.

Drasgow, F., & Kang, T. (1984). Statistical power of differential validity and differential prediction analyses for detecting measurement nonequivalence. Journal of Applied Psychology, 69, 498-508.

Dunham, R. B. (1974). Ability-skill relationships: An empirical explanation of change over time. Organizational Behavior and Human Performance, 12, 372-382.

Fleishman, E. A. (1953). A factor analysis of intra-task performance on two psychomoter tests. Psychometrika, 18, 45-55.

Fleishman, E. A. (1960). Abilities at different stages of practice in rotary pursuit performance. Journal of Experimental Psychology, 60, 162-171.

Fleishman, E. A., & Fruchter, B. (1960). Factor structure and predictability of successive stages of learning morse code. Journal of Applied Psychology, 44, 97-101.

Fleishman, E. A., & Hempel, Jr., W. E. (1955). The relation between abilities and improvement with practice in a visual discrimination reaction task. Journal of Experimental Psychology, 49, 5, 301-312.

Fleishman, E. A., & Hempel, Jr., W. E. (1954). Changes in factor structure of a complex psychomoter test as a function of practice. Psychometrika, 19, 239-252.

Fleishman, E. A., & Rich, S. (1963). Role of kinesthetic spacial-visual abilities in perceptional-motor learning. Journal of Experimental Psychology, 66, 6-11.

Ghiselli, E. E. (1956). Dimensional problems of criteria. Journal of Applied Psychology, 40, 1-4.

Gough, H. G., & Hall, W. B. (1975). The prediction of academic and clinical performance in medical school. Research in Higher Education, 3, 301-314.

Guion, R. M., & Gibson, W. M. (1988). Personnel selection and placement. Annual Review of Psychology, 39, 349-374.

Guttman, L. (1955). A generalized simplex for factor analysis. Psychometrica, 20, 173-192.

Henry, R. A., & Hulin, C. L. (1987). Stability of skilled performance across time: Some generalizations and limitations on utilities. Journal of Applied Psychology, 72, 457-462.

Henry, R. A., & Hulin, C. L. (1989). Changing Validities: Ability-performance relations and utilities. Journal of Applied Psychology, 74, 365-367.

Hinrichs, J. R. (1970). Ability correlates in learning a psychomotor task. Journal of Applied Psychology, 54, 56-64.

Hollander, E. P. (1965). Validity of peer nominations in predicting a distant performance criterion. Journal of Applied Psychology, 49, 434-438.

Hulin, C. L., & Humphreys, L. G. (1980). Foundations of test theory: Construct validity in psychological measurement. Proceedings of the Symposium on Theory and Application in Education and Employment (pp. 5-12). Princeton, NJ: U.S. Office of Personnel Management and Educational Testing Service.

Humphreys, L. G. (1968). The fleeting nature of the prediction of college academic success. Journal of Educational Psychology, 59, 375-380.

Humphreys, L. G. (1979). The construct of general intelligence. Intelligence, 3, 105-120.

Humphreys, L. G. (1960). Investigation of the simplex. Psychometrika, 25, 4, 313-323.

Humphreys, L. G. (1985). Intelligence: Three kinds of instability and their consequences for policy. Unpublished manuscript, University of Illinois at Urbana-Champaign.

Humphreys, L. G., & Davey, T. C. (1984). Intellectual growth from infancy to adulthood. Unpublished manuscript, University of Illinois.

Humphreys, L. G., & Parsons, C. K. (1979). A simplex process model for describing differences between cross-lagged correlations. Psychological Bulletin, 86, 325-334.

Humphreys, L. G., & Taber, T. (1973). Prediction study of the GRE and eight semesters of college grades. Journal of Educational Measurement, 10, 179-184.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). Meta-analysis. Beverly Hills, CA: Sage.

Jones, M. B. (1970). A two-process theory of individual differences in motor learning. Psychological Review, 77, 353-360.

Jones, M. B., Kennedy, R. S., & Bittner, Jr., A. C. (1981). A video game for performance testing. American Journal of Psychology, 94, 143-152.

Kaufman, H. G. (1972). Relations of ability and interest to currency of professional knowledge among engineers. Journal of Applied Psychology, 56, 495-499.

Lin, P., & Humphreys, L. G. (1977). Predictions of academic performance in graduate and professional school. Applied Psychological Measurement, 1, 249-257.

Noble, C. E. (1970). Acquisition of pursuit tracking skill under extended training as a joint function of sex and initial ability. Journal of Experimental Psychology, 86, 360-373.

Parker, Jr., J. F., & Fleishman, E. A. (1960). Ability factors and component performance measures as predictors of complex tracking behavior. Psychological Monographs, 74, Whole No. 503, 1-17.

Paulsen, G. B. (1935). The reliability and consistency of individual differences in motor control, Part I. Journal of Applied Psychology, 19, 29-42.

Paulsen, G. B. (1935). The reliability and consistency of individual differences in motor control, Part II. Journal of Applied Psychology, 19, 166-179.

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. Journal of Applied Psychology, 65, 373-406.

Powers, D. E. (1982). Long-term predictive and construct validity of two traditional predictors of law school performance. Journal of Educational Psychology, 74, 568-576.

Reynolds, B. (1952). The effect of learning on the predictability of psychomotor performance. Journal of Experimental Psychology, 44, 189-198.

Roff, M. (1941). A statistical study of the development of intelligence test performance. Journal of Psychology, 11, 371-386.

Rosenthal, R. (1979). The "file-drawer problem" and tolerance for null results. Psychological Bulletin, 86, 638-641.

Rothe, H. F. (1946a). Output rates among butter wrappers: I. Work curves and their stability. Journal of Applied Psychology, 30, 199-211.

Rothe, H. F. (1946b). Output rates among butter wrappers: II. Frequency distributions and an hypothesis regarding the restriction of output. Journal of Applied Psychology, 30, 320-327.

Rothe, H. F. (1947). Output rates among machine operators: I. Distributions and their reliability. Journal of Applied Psychology, 31, 484-489.

Rothe, H. F. (1951). Output rates among chocolate dippers. Journal of Applied Psychology, 35, 94-97.

Rothe, H. F. (1970). Output rates among welders: Productivity and consistency following removal of a financial incentive system. Journal of Applied Psychology, 54, 549-551.

Rothe, H. F. (1978). Output rates among industrial employees. Journal of Applied Psychology, 63, 40-46.

Rothe, H. F., & Nye, C. T. (1958). Output rates among coil winders. Journal of Applied Psychology, 42, 182-186.

Rothe, H. F., & Nye, C. T. (1959). Output rates among machine operators: II. Consistency related to methods of pay. Journal of Applied Psychology, 43, 417-420.

Schmidt, F. L, Gast-Rosenberg, I., & Hunter, J. E. (1980). Validity generalization results for computer programers. Journal of Applied Psychology, 65, 643-661.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.

Schmidt, F. L., & Hunter, J. E. (1981). Employment testing: Old theories and new research findings. American Psychologist, 36, 1128-1137.

Schmidt, F. L., Hunter, J. E., & Caplan, J. R. (1981). Validity generalization results for two jobs in the petroleum industry. Journal of Applied Psychology, 66, 261-273.

Schneider, B. J., & Bartlett, C. J. (1968). Individual differences and organizational climate. I. The research plan and questionnaire development. Personnel Psychology, 21, 323-324.

Stelmach, G. E. (1969). Individual differences and intra-individual variability in motor performance under continuous-practice conditions. Human Factors, 11, 201-206.

Thurstone, L. L. (1938). Primary mental abilities. Chicago: University of Chicago Press.

Vernon, P. E. (1950). The structure of human abilities. New York: Wiley.

Viteles, M. S. (1933). The influence of training on motor test performance. Journal of Experimental Psychology, 16, 556-564.

Wesman, A. G. (1956). Aptitude, intelligence, and achievement. Test Service Bulletin (No. 51). New York: Psychological Corporation.

Wilson, R. S. (1983). The Loisville twin study: Development synchronies in behavior. Child Development, 54, 298-316.

Winterbottom, J.A., Pitcher, B., & Miller, P.V. (1963). Report on the readministration of the LSAT to third-year students. Law School Admissions Council Reports of LSAC Sponsored Research, 1, 1949-69, Princeton, NJ: LSAC, 1976.